Preliminary Portfolio Analysis of a Cross-cutting Science Area using a Supervised Learning Approach:

NIGMS Technology Research and Development

Amy Swain, Kelley Smith, Paula Flicker, Stefan Maas, Pamela Marino, Ward Smith, Mary Ann Wu, Bin Zhou, Calvin Johnson

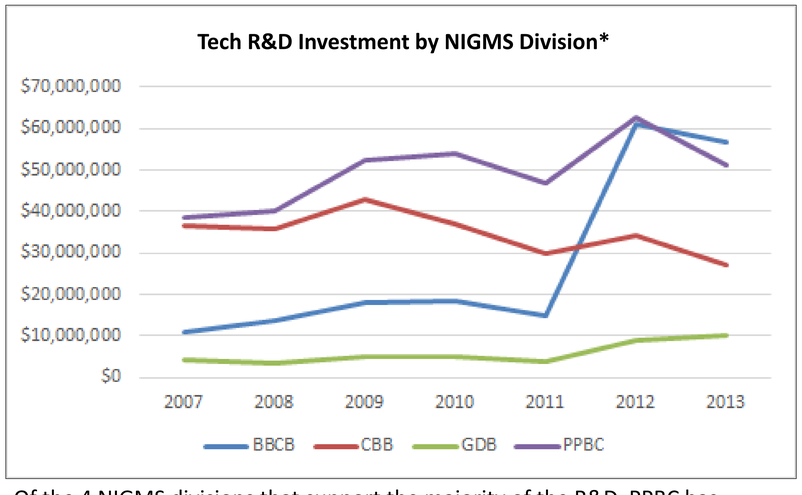
National Institute of General Medical Sciences

Background

Known as the "basic science" institute at NIH, the National Institute of General Medical Sciences (NIGMS) has always viewed technology research and development as an important component of its mission. As we look forward in the NIGMS strategic planning process, analyzing our historical investment in Technology R&D (Tech R&D) has become an important exercise that can inform our decisions about how, and in what areas, to invest in the future. There are substantial challenges associated with doing a portfolio analysis of the broad Tech R&D investment because it is a cross-cutting area that transcends the science supported by NIH. There is no 'Research, Condition, and Disease Categorization' (RCDC) category associated with Tech R&D, and approaches such as NIH Maps that group grants according to domain science area are ineffective for such a cross-cutting area. Therefore, in order to classify NIGMS grants as Tech R&D or not, and define the sub-categories within the area, we elected to use a supervised learning approach developed and executed by the NIH CIT Division of Computational Bioscience.

Introduction

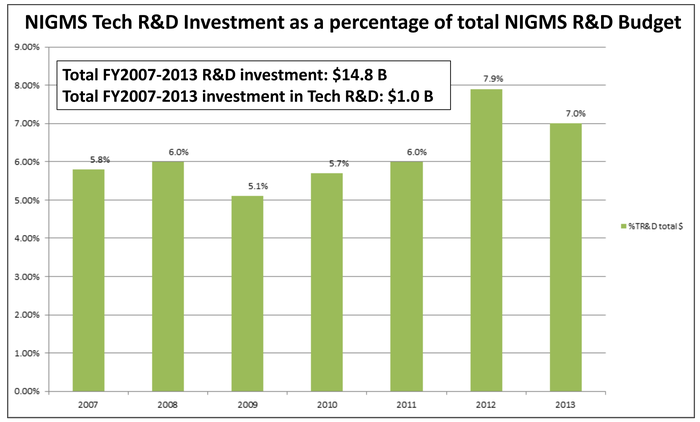
The **Input** data is a listing of NIGMS grants of which the abstracts and specific aims text is used for calibrating the model. In this approach, ensemble models are trained, using a Sampled, Augmented Ensemble Support Vector Machine (SAE-SVM) algorithm, from a manually generated and annotated list of positive (Tech R&D) and negative (not Tech R&D) research grants as defined by a committee of subject matter experts. Each of the trained SAE-SVM models are then used to make predictions about whether or not an NIGMS R&D grant awarded during Fiscal Year 2007-2013 should be categorized as Tech R&D. After this, the predictions from all the ensemble members are aggregated to generate an **Output** list of grants ranked in order of degree to which they match the positive Tech R&D training set. Using this approach, we have identified a set of NIGMS grants that have contributed the most to support of technology research and development between 2007 and 2013. We present an analysis of this preliminary data set, which contains 3021 grants.



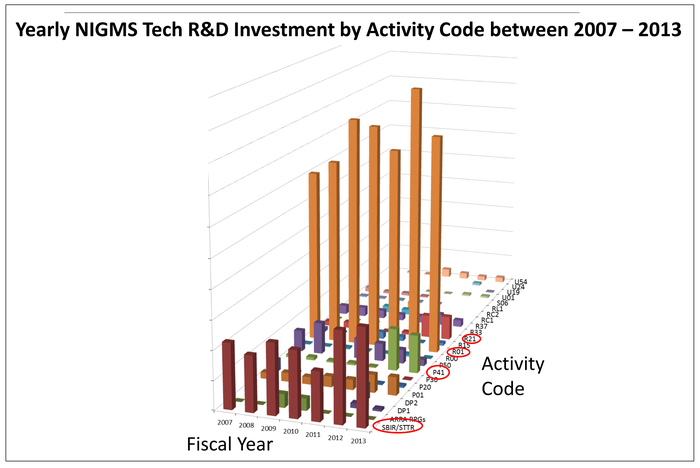
Of the 4 NIGMS divisions that support the majority of the R&D, PPBC has supported most of the TR&D. This changed in 2012, when some NCRR+ programs joined NIGMS, with the majority of Tech R&D moving to BBCB.

*NIGMS Divisions that support most of the R&D
BBCB – Biomedical Technology, Bioinformatics and Computational Biology
CBB – Cell Biology and Biophysics
GDB – Genetics and Developmental Biology
PPBC – Pharmacology, Physiology and Biological Chemistry

+National Center for Research Resources



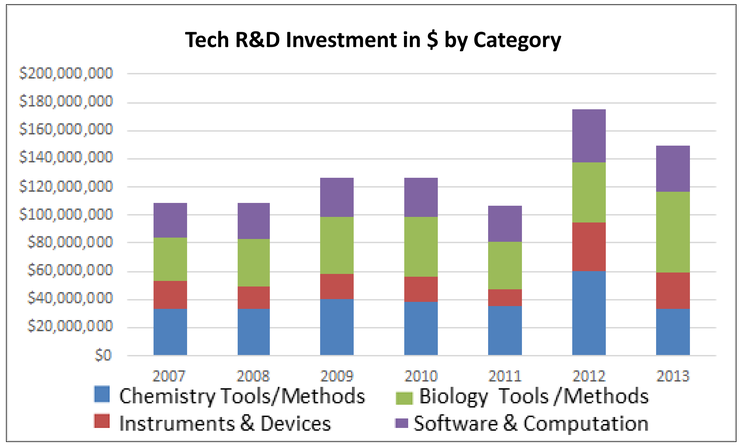
Over the FY2007-2013 period, Technology R&D support has comprised 6.7% of the NIGMS R&D budget.



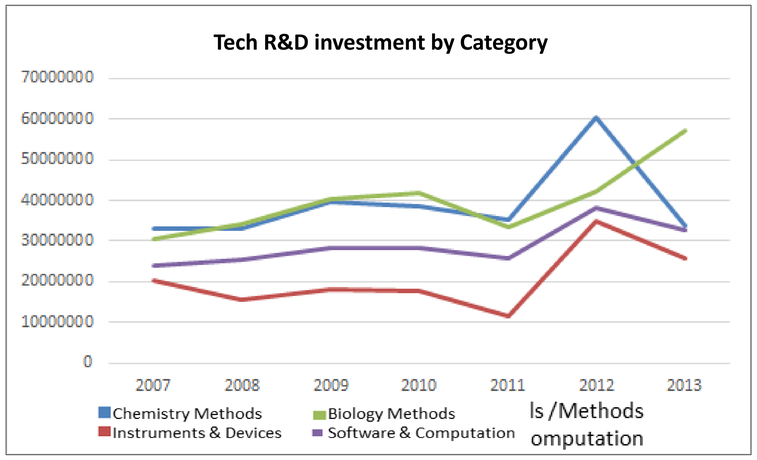
The greatest Tech R&D investment used the R01 activity code, followed by SBIR/STTR. Slightly above the background of most other grants in 2007-2011 were P50 awards. After integration of the NCRR Biomedical Technology R&D portfolio in 2012 and 2013, investment in Technology R&D using the P41 and R21 activity codes is apparent.

Methods – Each cycle of training and running the algorithm is noted as 'SVM'. '**Positive**' means grants that were identified as Tech R&D grants.

	SVM 1	SVM 2		SVM 3	SVM 4
Input Source	Training set generated by Program Officers from their division's portfolio. Technology development was major emphasis for award	SVM1 grants from 2012-2013 reviewed by Program Officers and designated Positive or Negative. Each grant reviewed by 2 Program Officers.	Because we found that the Output lists were not sufficiently discriminating between grants that were Positive and Negative for Tech R&D, we defined four subcategories, listed below.	SVM 2 input grants divided into categories based on comments from Program Officers	SVM3 retrievals were examined by Program Officers and designated Positive or Negative. Each grant reviewed by 2 Program Officers
	Positive*	**Positive**	**Category**	**Positive**	**Positive**
Input - list of grants used for training the model	212	312	Instrumentation & Devices	97	175
			Software & Computation	48	135
			Methods&Tool Development	82	236
			Chemistry Methods&Tool Dev.	8	153
Output - list of retrieved grants above a threshold	858	666	Instrumentation & Devices	222 (133 new)	218 (84 new)
			Software & Computation	239 (209 new)	201 (90 new)
			Methods&Tool Development	265 (201 new)	336 (150 new)
			Chemistry Methods&Tool Dev.	210 (200 new)	357(218 new



The total investment was highest in 2012.



The Chemistry and Biology Tools/Methods categories are at similar levels until 2012 when Chemistry Tools/Methods investment spiked. It then declined in 2013 while Biology Tools/Methods investment increased.

Conclusions

- For portfolio analysis of a broad, inclusive research and development area, Supervised Learning was an effective way to identify grants in that area.
- Sub-categorization of the area of interest yielded improved results in terms of precision and recall, i.e. retrieving the largest number of Positive grants without including many that were not (False Positives).
- A team rating approach was effective for minimizing bias, and for sharing the substantial effort involved in rating multiple iterations.
- With further iteration, we expect to improve our data set in terms of precision and recall.
- Analysis **of the final data set** will inform future decisions for NIGMS investment in Tech R&D.